# A Vector Worth a Thousand Counts

*A Temporal Semantic Similarity Approach to Patent Impact Prediction*

Hain, Daniel; Jurowetzki, Roman; Buchmann, Tobias; Wolf, Patrick

*Publication date:*
2022

*Document Version*
Early version, also known as pre-print

# A Vector Worth a Thousand Counts*

## – A Temporal Semantic Similarity Approach to Patent Impact Prediction –

Daniel S. Hain[†] [ϕ], Roman Jurowetzki[ϕ], Tobias Buchmann[ψ], and Patrick Wolf[ψ]

[ϕ] *Aalborg University, Department of Business and Management, IKE / DRUID, Denmark*
[ψ] *Centre for Solar Energy and Hydrogen Research Baden-Württemberg (ZSW)*

**Abstract:** Patent data has long been used as a widely accessible measure of the rate and direction of technological change. However, this long tradition of research has so far focused on producing and analyzing measures of patent quantity, assuming the number of patents produced to accurately capture the rate of progress and innovation output. Existing attempts to measure patent quality are mostly limited to the use of forward- and backward citation pattern. In contrast, in this paper, we derive a patent quality indicator by leveraging the rich but up to now under-utilized textual information in patent abstracts. We employ vector space modeling techniques to create a high-dimensional vector representation of the patents to capture their technological signature. Using almost near linear-scaling approximate nearest neighbor matching techniques, we are able to compute dyadic similarity scores across large bodies of patent data. Based on the temporal distribution of a patents similarity scores, we compute ex-ante indicators of a patentÂťs technological novelty and ex-post indicators of technological impact and significance. At the case of circa 132.000 electro-mobility patents, we demonstrate the proposed indicators‘ to map, analyze, and predict patent quality on individual, firm, and country level, and its development over time.

**Keywords:** Technological change, patent data, natural-language processing, vector space modeling, quality indicators

*We would like to acknowledge the generous computational resource contribution by Google Cloud Platform;
Financial support for ZSW's research by BMBF Kopernikus ENavi (FKZ:03SFK4W0)
[†]Corresponding author: dsh@business.aau.dk

# 1 Introduction

> *"We have the choice of using patent statistics cautiously and learning what we can from them, or not using them and learning nothing about what they alone can teach us"*
>
> — *(Schmookler, 1966, p.56)*

Understanding the pattern and drivers of technological change is a crucial precondition to formulate meaningful long-term research and industry policy, firm level strategic decisions, and lucrative investment plans. Patent data here has long been used as a widely accessible measure of inventive and innovative activity and performance. Yet, it has widely been recognized that the technological as well as economic significance of patents varies broadly (Basberg, 1987). The long tradition of research utilizing patent data has so far mostly focused on producing and analyzing measures of patent quantity, assuming the amount of patents produced by a certain country, firm or within a technology class to capture the rate of progress and innovation output of the entity under investigation. Existing approaches to derive more nuanced indicators of patent quality are mostly limited to the analysis of the number or composition of a patent's International Patent Classification (IPC) assignments (Lerner, 1994), backward (Lanjouw and Schankerman, 2001; Shane, 2001; Trajtenberg et al., 1997) and forward citations (Harhoff et al., 2003a; Trajtenberg et al., 1997)

In this paper we attempt to derive a quality measure on patent level capturing its *ex ante* level of novelty as well as its *ex post* future technological impact by exploiting the rich textual information contained in the patent abstract. We employ vector space modeling techniques to create a high-dimensional vector representation of the patents under study. We evaluate the information-richness contained in this signature vectors by using them to predict the patents' IPC classes. We use fast approximate nearest neighbor matching to identify the most similar patents, for which we compute similarity scores with the focal one. We thereby avoid the step of computing a full similarity matrix and instead benefit from the near linear scaling of this method, making it appropriate for processing massive data sets. While the similarity scores are useful

in their own rights, for instance to analyze patent-similarity networks, we proceed in exploiting the temporal distribution of similar patents. Composing the average similarity score to earlier patents, we compute an *ex ante* indicator of *novelty*, and likewise for later patents we *ex post* measure their *impact* and technological significance. We further evaluate our indicator against a variety of common measures of patent quality. At the case of circa 132.000 electro-mobility related patents we demonstrate the usefulness of this measure to map, analyze, and predict patent quality on individual, firm, and country level, and its development over time.

We thereby contribute to the existing body of literature in several ways. First, we demonstrate the usefulness of semantic-based approaches to measure novelty and the impact of patents as well as to map relational structures between them. Second, we provide a method pipeline capable of producing semantic similarity scores for massive amounts of data. Third, in our case analysis we shed light o9n the development of the technological life-cycles, and the catching up of latecomer countries entering the technological field.

The remainder of the paper is structured as follows. In section 2, we review literature on the relationship between patents and the rate and direction of technological change, innovation output and performance on firm as well as country level. We further review and discuss research utilizing patent data to map technological change, and derive measures of patent quality. In section 3, we discuss methodological considerations and describe our approach to create a semantic-based measure of patent quality. We apply this set of methods in the following section 4 at the case of electro-mobility, demonstrate its usefulness by providing insights in the sector's development, and derive quality measures on different levels of aggregation. In section 5, we review our set of methods and the results produced, provide suggestions for further applications, extensions, and avenues for future research. Finally, section 6 concludes.

# 2 Theoretical Considerations and Literature Review

> *"Ideally, we might hope that patent statistics would provide a measure of the output of inventive activity, a direct reading on the rate at which the potential production possibilities frontier is shifted outward. The reality, however, is very far from it. The dream of getting hold of an output indicator of inventive activity is one of the strong motivating forces for economic research in this area"*
>
> — *(Griliches, 1990, p.1669)*

## 2.1 Patents as a measure invention, the rate and direction of technological change

A wide body of literature in economics and other areas of the wider literature on innovation studies has long embraced patents as a measure of the rate as well as direction of technological change. Indeed, the correlation between the number of patent applications, and various measures of innovation output and success have been empirically investigated and established at various levels, such as countries, sectors, and industries (Pavitt, 1985, 1988).

However, the meaningfulness of patents to map the pattern as well as measure the rate of technological change is also perceived to be limited by the fact that: (i.) not all inventions are patentable, (ii.) not all patentable inventions are patented, (iii.) not everything patented represents an invention, (iv.) the importance of patents as a mean of intellectual property protection varies broadly across jurisdictions, industries, and over time (Pavitt, 1985, 1988). Consequently, any attempt to capture inventive activity by drawing from patent data will be subject to systemic false negatives as well as false positives. Still, as (p.22 Niosi, 2005) suggests: "Even if not all commercially useful novelties are patented, not all patents are exploited in the market, and the exploitation may occur in a place different from the one where the innovation took place, no other indicator is better suited to the study of innovation." Indeed, given all it's flaws, no other indicator of inventive activity up to now is widely accessible across firms, industries, countries, and over time. Further, patent data can be considered

information-rich, since it captures individuals (inventors), their corporate association (applicants), proxies of knowledge the patent is based on (citations), and detailed descriptions of their content (abstract).

## 2.2 Patents as a measure of innovation performance

On the firm level, a strong relationship between patent numbers and R&D expenditures has been found, implying that patents are a good indicator of differences in inventive activity across firms (Griliches, 1990). Besides its use as an indicator of inventive activity, previous research shows that patents are a valid indicator for the output, value and utility of inventions (Trajtenberg, 1990). Moreover, there is a relatively large body of literature which uses patents not only as a proxy for inventions but also for innovations (Hagedoorn and Cloodt, 2003) as well as resulting economic performance on firm level (Ernst, 2001). These signals are also useful and recognized by investors (Hirschey and Richardson, 2004), making them more likely to provide firms with external capital (Hall and Harhoff, 2012).

Yet, it has long been recognized that the technological as well as economic significance of patents varies broadly (Basberg, 1987). While all patents must meet objective criteria in terms of novelty and utility in order to be granted, this can still be an incremental and narrow improvement to an existing technology, invisible in its impact on technological progress. Even when radically novel and theoretically of large technological scope and broadly applicable, its economic value is contingent to firm, technology, market, and timing related factors.

## 2.3 Patent quality measures

Generally, patents are considered as a useful indicator of inventive activity, and albeit heterogeneous still indicative of technological and economical significance and value. To strengthen this association, a large body of literature has explored the rich information contained in patent data to construct patent quality measures.[1]

[1]For a recent and exhaustive review on patent quality measures, consider Squicciarini et al. (2013)

Such measures are mostly derived from information on a patent's (i.) composition of IPC classes, (ii.) the number and pattern of backward citations or (iii.) forward citations and (iv.) strategic reaction by applicant and competitor firms on the patent.[2] While the former two are *ex-ante* measures readily available at the point a patent is granted, the latter are cumulative over time, leading to a truncation that makes such measures only *ex-post* after a sufficient period of time meaningful.

*Ex-ante* indicators mainly aim at deriving patent quality by the scope and novelty of the patent's underlying technology. This has been approximated by the patents technological *scope* (Lerner, 1994) in terms of the number of IPC classes the patent is assigned to. The intuition is that the more different IPC classes a patent covers, the larger the technology's generality and potential usefulness across applications and industries. In addition, backward citations are commonly used as a measure of novelty in terms of applied and combined knowledge. While it has been argued that the number of backward citations to capture the amount and scope of applied knowledge and assign it a positive association with radical invention (Schoenmakers and Duysters, 2010) and patents value (Harhoff et al., 2003b), large numbers of backward citations might also indicate a more incremental invention (Lanjouw and Schankerman, 2001). In that sense, having no backward citations is seen as a potential sign of radically new inventions breaking away from established technological trajectories (Ahuja and Lampert, 2001). In addition, the citation of non-patent literature (NPL, mainly academic papers) (Narin et al., 1997) shows closeness to scientific knowledge (Tong and Davidson, 1994), and therefore might be more likely to be general and radical. In a more nuanced way, the composition of backward citations is used to provide information of the structure how technologies are combined and applied in the patent. Popular examples are the breadth of cited IPC classes (*originality* index, Trajtenberg et al., 1997), or the number of cited IPC classes different from the ones of the citing patent (*radicalness* index, Shane, 2001).

[2]Due to the large body of literature on that issue, such a list can for the sake of brevity not be exhaustive. Further common measures not discussed in detail here include the size of the patent family (Harhoff et al., 2003b) and the lag between the patent application and approval Harhoff and Wagner (2009),

*Ex-post* patent quality indicators aim at capturing the patent's economic or technological impact. Again, while such retro-perspective indicators have the potential to measure real and not only potential impact, this comes at the expense of often long delays between the patent application and the possibility to create meaningful measures. Economic impact is mostly measured indirectly by firm action which might reveal the perceived value of the patent, such as the number of claims filed on the patent (Tong and Davidson, 1994), and the renewal of a patent after its expiration (Pakes and Schankerman, 1984). Technological impact is to the largest extend approximated by the number or pattern of the patent's received forward citations. First, the raw number of citations received can be interpreted as a measure of general technological usefulness, which indeed has been shown to correlate well with other quality measures such as its technological and economic value (Harhoff et al., 2003a), its contribution to firm market value (Hall et al., 2005) and the inventor assessments of its economic value (Gambardella et al., 2008). Related, some scholars focus on the study of *breakthrough invention*, defined as the top 1% cited patents in a certain field (Ahuja and Lampert, 2001). To utilize not only the amount but also the structure and composition of forward citations, approaches similar to the ones on backward citations have been developed, such as the *generality* index (Trajtenberg et al., 1997) which measures the range of technological fields that cite the patent.

Lately, an increased effort has been made to develop composite indicators of patent quality, which incorporate different *ex-ante* as well as *ex-post* concepts (eg., Dahlin and Behrens, 2005; Lanjouw and Schankerman, 2004; Verhoeven et al., 2016).

## 2.4 Natural language based approaches to create patent measures

Moving beyond simple patent counts, patent citation analysis and the development of patent novelty, quality, and impact indicators has proven to deliver valuable new insights into technology development paths. Still, the main drawback of limited scope of information by ignoring technology descriptions cannot be solved without performing a semantic analysis of patent abstracts and/or bodies.

As a first step towards a more complete analysis of patent documents keyword-based methods have been introduced (Lee et al., 2009). These studies are based on the comparison of the occurrence of keywords measuring the term frequency by classifying keywords. However, working with isolated keywords is often too unspecific to learn something about the deeper meaning of a patent. Hence, multi-word analysis to better reflect the technological content of a patent has been introduced (Gerken and Moehrle, 2012). Going further, Gerken and Moehrle (2012) measure patent novelty by generating similarity matrices to highlight those aspects of an invention that have already been delineated in previous patents. That is, the statistical comparison of the semantic structures of patents within a given patent set results in a matrix of similarity values. Passing and Moehrle (2015) measure technological convergence in the field of smart grids by a semantic patent analysis approach.

In their literature review on the state-of-the-art in patent analysis Abbas et al. (2014) define mining techniques as a knowledge based process that deals with extracting useful patterns from unstructured data rather than structured data. The text mining techniques used in patent analysis are typically based on natural language processing (NPL), neural networks based approaches and semantic analyses. NLP is based on text mining performed with computational techniques to (i) represent textual information of documents and (ii) to analyze this information. Such approaches can broadly be categorized into previously mentioned keyword-based studies and more advanced Subject Action Object (SAO) based studies. A major advantage of the SAO approach is that it can analyze unstructured information by representing the relationships among key technological components. NLP based approaches are particularly effective in processing large documents that contain rich textual data. Therefore, semantic approaches have been used to study a variety of aspects of patent information: Kim et al. (2016) use semantic patent topic analysis to generate patent development maps to identify technological trends. Tran and Kavuluru (2017) study the problem of patent classification with respect to the new Cooperative Patent Classification (CPC) system. They consider patent classification as a multi-label text classification problem.

The authors propose a supervised classification system that exploits the hierarchical taxonomy of CPC as well as the citations of a patent. To evaluate their approach, they conducted experiments on US patents released in 2010 and 2011 for over 600 labels. In a similar vein, Fall et al. (2003) study patent classification for IPC classes by applying different machine learning techniques such as support vector machines (SVM), naive Bayes and k-nearest neighbors (k-NN) where they find that SVM is superior to other methods. Liu and Shih (2011) apply a hybrid system for USPC classification that is based on patent network analysis as a complement to a text based analysis. Don and Min (2016) study feature selection for automatic categorization of patent documents. In their study they propose an algorithm that automatically categorizes patent documents by considering the structural information of the patents. Among others, they apply feature selection based on term frequencies. Feature vectors are constructed from the structural information of the patent. Classifications are conducted using a random forest (RF), SVM and Naive Bayes (NB) classifiers. It was found that the semantic structural information of a patent document is an important feature set for constructing the terms of a document for classification.

## 3 A Semantic Approach to Measuring Patent Quality

Our aim is to create a measure of "future orientation" for every patent in our dataset, which we interpret as the potential impact of the technology embedded in the patent and expressed by the semantic properties of its description. To do so, we first have to assume that every patent can be represented as a vector $v$ in some vector space $V \in \mathbb{R}^n$ such that vectors satisfy two properties: *composability* and *comparability*. Vectors must be composable so that we can compute a signature vector for every patent, which can be manipulated using vector algebra, for instance to compute an average vector for an aggregated higher level entity such as a firm or country. In addition, such vectors need to be comparable, so that for any pair of vector $\overrightarrow{i}$ and $\overrightarrow{j}$, a robust similarity score $s(\overrightarrow{i}, \overrightarrow{j})$ can be computed. If such a vector indeed represents the technological properties of a patent accurately, the resulting similarity score $S_{i,j}$ provides a dyadic measure

of technological relatedness, which can be used for static mapping but also dynamic analysis. Our measure of patent quality mainly exploits the temporal distribution of top-similarity scores. In the following, we describe the techniques, parameters, and general logic behind every step of the patent quality measure computation in detail. Figure 7 depicts the preprocessing pipeline that we use to identify similar patents in the corpus and to calculate dyadic relatedness measures.

### 3.0.1 From patent to vector: Vector space modeling

Given a relatively high number of patent abstracts, we required an efficient approach to generating numeric representations of the patent text that would preserve its semantic features. There are several approaches to doing this with a rapid development of new methods in the recent years. The most basic approach would be to represent individual abstracts as word-co-occurrence vectors, i.e. an array of dummies, or weighted for generality and specificity of the utilized terms (TFIDF). Even when stemming the terms and excluding rare expressions as well as stop words and other terms that only contribute little to the meaning structure of the document, that would usually lead to extremely sparse and high-dimensional representations. Therefore recently researchers started utilizing more advanced dimensionality reduction techniques such as latent semantic indexing (LSI, Deerwester et al. (1990); Dumais et al. (1988)) and topic modeling techniques e.g. Latent Dirichlet allocation (LDA, Blei et al. (2003)) that not only decrease complexity but to some extent for relational attributes across the analyzed corpus. Such models, once trained on an appropriate corpus, allow to represent a new document as a vector of a predefined dimensionality.

More recently word embedding approaches e.g. Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) have gained traction. Here the model learns term meanings from the context that surrounds the term rather than merely within-document co-occurrence. Training of such models on large datasets allows to account for syntax and to extract higher level meaning structures for terms. Summing and averaging such word vectors, has proven to generate excellent document representations.

On a high level, the most recent approaches in Natural Language Processing combine elements from both traditions, drawing on the current developments in deep learning. For this project, we use the *spaCy* (Honnibal and Johnson, 2015) Python library to generate abstract vector representations. SpaCy is a performance optimized library that does not require preprocessing of the input data and is thereby very flexible. In this version of the paper we use the *en_core_web_lg* model that is an English multi-task convolutional neural network model trained on OntoNotes, with GloVe vectors (685k unique vectors with 300 dimensions) trained on Common Crawl.

Given a patent abstract, *spaCy* predicts the meaning of each term in the document based on matching word vectors, word-concurrence and syntax. Thus word vectors are predicted based on the context also for terms that the model did not encounter before. The combined vector of an abstract is the average 300-dimensional vector of all contained terms. Since the individual word vectors are syntax dependent, the average vector is more than the sum of all elements but represents the meaning structure of the patent abstract.

To access the reliability of the vector representations, we use a predictive approach, aiming at a correct prediction of the (IPC) classification symbol on class level, assuming that the main IPC symbol encodes a high level meaning of the text. For that we train a deep artificial neural network [3] splitting the dataset 176.061 patents (13 individual classes) where the symbols appear at least 1000 times across the corpus into a training set (80%) and a test set (20%). Marginal improvements of the prediction accuracy are minimal after 20 epochs with a batch size of 500. Overall the model achieves an average prediction accuracy of 73%. To compare with other model classes, we also evaluated the performance of *spaCy* vectors against LSI generated vectors on a subset of 10.000. Results were comparable. Yet, future versions of this project will aim at integrating domain adaptation of the general model to better account for the technical vocabulary used in patent abstracts. Since, we are overall mainly interested in defining a patent

[3] 5 deep layers (512,1024,512,128,64), 5 dropout layers (rate = 0.3 to 0.1) for overrfitting prevention, 1.333.517 trainable parameters. Implemented using the Keras library with TensorFlow backend on a Google Colab GPU

quality indicator from relational attributes between patents (semantic similarity) and our model is equally applied to all abstracts, that are assumed to use a normalized technical language, the large multi-purpose model that we selected should suffice.

## 3.1 From vector to similarity: Approximate nearest neighborhood matching

We are interested in identifying $k$ most similar documents in the corpus. The most precise but also naive approach is a brute-force nearest neighbor search where a similarity score (e.g. euclidean distance) for each pair of observation is calculated for instance by taking the dot product of the document matrix and its transponse. For $N$ samples in $D$ dimensions this approach scales as $O[DN^2]$, in our case with $D = 300$ and $N \approx 132.000$. While such an approach is competitive for small samples, it becomes infeasible as the sample and thereby the complexity $O$ grows. In our case alone the resulting $NN$ matrix would have to store $1.73e^{10}$ values, making this approach not feasible.

Efficient nearest neighbors computation is an active area of research in machine learning and one of the common approaches to this problem is using k-d trees that partition the space to reduce the required number of distance calculations. Search of nearest neighbors is then performed by traversing the resulting tree structure. Utilizing such an approach can reduce complexity to $O[DNlog(N)]$ and more. In our case this would leads to an efficiency increase by a factor of at least $1.12e^4$.

We utilize the efficient *annoy* (Approximate Nearest Neighbor Oh Yeah!, Bernhardsson (2017)) implementation that constructs a forest of trees (100) using random projections. We then retrieve the 200 nearest neighbors along with the calculated euclidean distance.

### 3.1.1 From similarity to patent quality

Our resulting similarity index between patents based on the semantic of the patent abstracts appears valuable on its own right, since it offers a nuanced measure of relat-

edness which is in contrast to citations not dependent on explicit mentioning by the author or patent office. As a dyadic measure, the derived semantic similarity can also be used to create patent networks, as we demonstrate later. Such a relational representation offers the potential to visually map technological fields and their development, derive further network related measures such as degree centrality, betweenness, and perform relational clustering exercises.

However, to develop a measure of patent quality, we exploit the temporal properties of our similarity measure. In contrast the commonly used patent citations, semantic similarity includes relationships to other patents independent of the time of application. Therefore, for every patent $i$, the set of mostly semantically similar patents $J_i[1 : m]$ will contain patents $j$ with earlier as well as later application dates. With that information, we construct a temporal similarity index on patent level as follows:

$$q_i = \sum_{j=1}^{m} \frac{\{\Delta t_{j,i} > \tau\}\ s_{i,j}}{m} \tag{1}$$

Consequently, $q_i$ represents patent $i$'s share of similar patents with application date in the future, weighted by their similarity $s_{i,j}$. The parameter $\tau$ represents the time delay after which a patent $j$ is considered to be in the future. To offset the delay between patent application and the official publication of 6 to 12 months (Squicciarini et al., 2013) we set $\tau = 1$, meaning that patents with application date more that a year after the focal patent are considered as laying in the future.

### 3.2 Context and data

The patent data we used for our study was retrieved from the EPO's Patstat (spring 2017 edition) worldwide patent database which covers bibliographic patent data from more than 100 patent offices over a period of several decades. To identify the relevant patents, we make use of the international patent classification (IPC) codes. The IPC system has been introduced to obtain an internationally uniform classification of patent documents. In particular, the classification serves as (a) an instrument to facilitate access to the technological and legal information contained in patent documents; (b)

a starting point for selective dissemination of patent information; (c) a starting point for scrutinizing the state of the art in a specific field of technology; (d) a source for industrial property statistics. The classification system becomes periodically revised to improve it and consider recent technological developments. The IPC system is hierarchical structured. It is divided into the following levels: section, class, subclass, group and forms altogether the complete classification symbol. Sections are the highest level of hierarchy of the classification. Each section is subdivided into classes which are the second hierarchical level. Each class consists of one or more subclasses and each subclass symbol consists of the class symbol followed by a capital letter. We predict classifications at the class / subclass level. Classification is typically performed at the class or subclass level. This relates to the observation that the labels at the subclass level are more static whereas group and subgroup labels are revised more often (WIPO, 2017).

### 3.2.1 Context: Electro-mobility technologies

Electric vehicles (EVs) are relatively energy efficient that is for instance why the German governments seeks to increase the number of EVs by the year 2030 to 6 million. Their efficiency rate exceeds 90% while the internal combustion engines reaches just about 35%. This has for instance to do with the possibility to recuperate braking energy back into the vehicle's energy supply system. In contrast, with the combustion engine it would be lost, converted into heat. Moreover, EVs produce no direct emissions. The air pollution problem is shifted from roads to the energy sector. Once a renewable energy system has been established, that is based on solar energy, wind energy etc., the overall emissions can vastly be reduced and electric vehicles will be environmental friendly (Larminie and Lowry, 2003). Apart from being more environmentally friendly, electric vehicles have a number of further advantages: "Electric motors are low-maintenance, versatile and exceptionally quiet" (Deffke, 2013, 4). "Electric vehicles" is a relatively broad concept and there are several different types of electric vehicles. Typically we can distinguish four types of EVs: battery electric vehicles

(BEV), hybrid electric vehicles (HEV), range extended electric vehicles (REEV) and fuel cell vehicles (Proff and Kilian, 2012). The engine system unit of a BEV consists of the following main parts: battery, electric motor and controller. The battery system is being charged with the special power mains. It stores the energy and supplies the electric motor. The controller is called so, because it controls the amount of electricity that the motor gets from the battery as well as the speed of the vehicle. Hybrid vehicles include both a combustion and an electric engine (Larminie and Lowry, 2003).

### 3.2.2 Patent data

The patent data used for our study originates from the EPO's Worldwide Patent Statistical Database (Patstat), which contains bibliographic data for patents of developed as well as developing countries. In order to identify relevant patents, we used the International Patent Classification (IPC). Our analysis focuses on a key technology of battery electric vehicles (BEVs) which is the electric propulsion. Pilkington et al. (2002) studied patent data as indicators of technological development related to electric vehicles. Representing a core IPC class of electric vehicle technology they use the class B60L11/- IPC which represents "Electric propulsion with power supplied within the vehicle". This is part of the broader IPC class B60 (vehicles in general) and the sub-class B60L (electric equipment or propulsion of electrically-propelled vehicles; magnetic suspension or levitation for vehicles; electrodynamic brake system for vehicles). However, we need to bear in mind that the class not only covers electric cars but also other electric vehicles such as for instance marine vehicles. Thus, for analysis the class B60L 11/00 and its subclasses were used, as they can be determined as a "likely home for EV patents" (Pilkington and Dyerson, 2006, 85). A list of all used IPC-classes and their description is given in Table 1.

For all patents related to the identified IPC-classes we extracted title and abstract as well as information about the publication year, the country, the IPC-class and the associated company. Further, all patents which cite the identified patents were collected including the same additional information, respectively. Due to quality and

comparability reasons, only filed patents containing complete information were used for analysis.

Table 1: List of used IPC-classes

| IPC class | Level | Description |
|---|---|---|
| B60L 11/00 | 0 | Electric propulsion with power supplied within the vehicle |
| B60L 11/02 | 1 | using engine-driven generators |
| B60L 11/04 | 2 | using dc generators and motors |
| B60L 11/06 | 2 | using ac generators and dc motors |
| B60L 11/08 | 2 | using ac generators and motors |
| B60L 11/10 | 2 | using dc generators and ac motors |
| B60L 11/12 | 2 | with additional electric power supply, e.g. accumulator |
| B60L 11/14 | 2 | with provision for direct mechanical propulsion |
| B60L 11/16 | 1 | using power stored mechanically, e.g. in flywheel |
| B60L 11/18 | 1 | using power supplied from primary cells, secondary cells, or fuel cells |

# 4 Analysis: First Insights

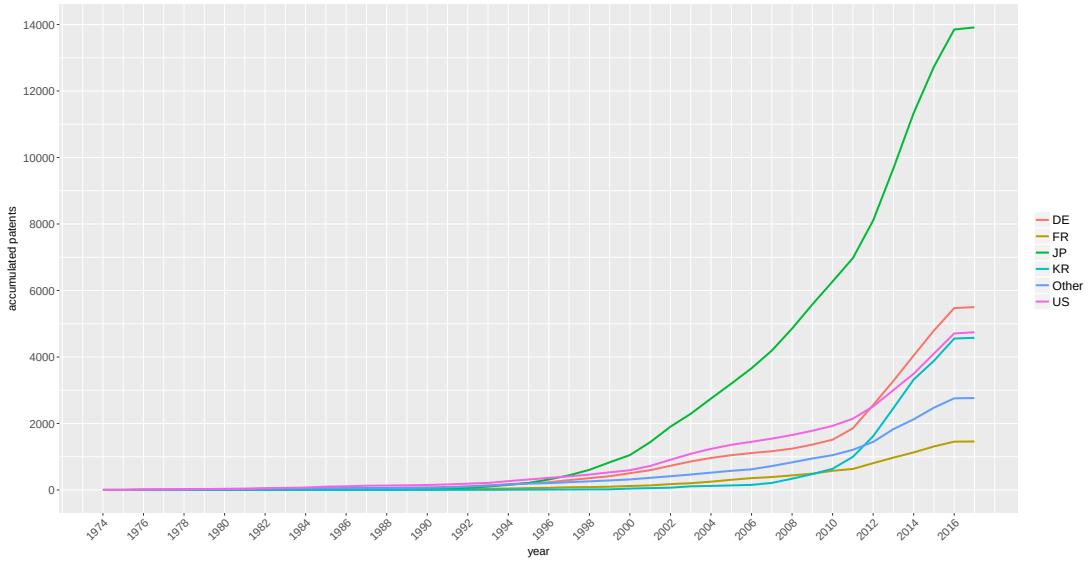
## 4.1 Descriptives

We start our analysis with a general overview of the collected data and its structure. Figure 1 shows the accumulated patent output of the five most productive countries (Japan, Germany, USA, South Korea and France) between the years 1974 and 2017. While the patent output for all shown countries remains on a constantly low level within the first two decades, a steady increase can be spotted since the second half of the 1990s.[4] Based on the displayed curves Japan can clearly be identified as the leading country in the field of core electric vehicle patents showing a sharp increase in output since the 2000s. The trends of Germany and the US are rather similar to each other but clearly remain behind the development of Japan. The patent output of South Korea and France remained comparatively low until a noticeable increase beginning in 2010. However, in the end only South Korea seemed to be capable of

[4] The noticeable stagnation starting in 2015 is caused by a lag in integration of new patents into the PATSTAT database and should therefore not be evaluated as an actual decline in patent output.

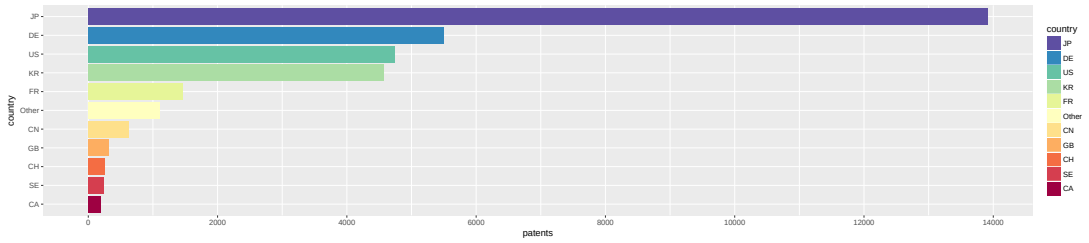
catching up with Germany and the US.

Figure 1: Cumulative patent output by country and year




An overview of the share of different countries on the total patent number is provided by figure 2. The chart clearly emphasizes the leadership of Japan, which can be assigned to 41% of all patents filed. In general, the data indicates a high output concentration on only few originators. Therefore, the leading five countries are the creators of 89% of all patents filed.

Figure 2: Share of patents by country




For our analysis, we extract all patent applications in the formerly described core IPC classes related to electro-mobility in from 1975 until 2017, resulting in a number of slightly above 60.000. To not be fully dependent on pre-defined IPC classes to capture patents related to our technological field under study, we further include all patents in

the same period which cite them but are assigned to other IPC classes (ca. 80.000), expanding our dataset to around 140.000 patents.

To upfront filter out patents irrelevant to the development of the technological field due to very low similarity to past as well as future developments, we remove all dyadic similarity scores below a value of 0.1, and then remove patents $i$ which match with less than 10 (of the initial 200) similar counterparts $j$. This almost leaves us with 76.990 remaining patents. While this step reduces our number of observations by almost 50%, it ensures that only patents with a minimal level of relatedness to the technological development of the field of electro-mobility are included in our analysis.

Table 2: Descriptive Statistics: Patents

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| $\Delta^t$ | 76,990 | 1.095 | 0.855 | −2.591 | 4.556 |
| $t^{past}$ | 76,990 | 0.209 | 0.100 | 0.000 | 0.813 |
| $t^{present}$ | 76,990 | 0.292 | 0.092 | 0.000 | 0.917 |
| $t^{future}$ | 76,990 | 0.499 | 0.135 | 0.043 | 1.000 |
| $sim$ | 76,990 | 0.557 | 0.067 | 0.124 | 0.781 |
| $sim^{past}$ | 76,990 | 0.116 | 0.059 | 0.000 | 0.476 |
| $sim^{present}$ | 76,990 | 0.165 | 0.058 | 0.000 | 0.600 |
| $sim^{future}$ | 76,990 | 0.276 | 0.081 | 0.026 | 0.656 |

Table 2 provides some first descriptive statistics on patent level. $\Delta^t$ indicates the average difference in years between the focal patent $i$ and the most similar patents $J[1, \ldots, m]$, which takes a value of around one. Over the whole life-cycle of a technology, one would expect $\Delta^t$ to have a mean close to zero, since patents overall tend to be most similar to the ones developed around the same time. In times of technological breakthroughs, this value should be higher than zero, and in times of technological exhaustion smaller. A value higher than zero therefore indicates that the technological field of electro-mobility has yet not reached full maturity. The variables $t^{past}$, $t^{present}$, and $t^{furure}$ indicate the share of the most similar patents with application date in the past ($-\Delta^t_{j,i} > \alpha$), present ($\pm\Delta^t_{j,i} \leq \alpha$), or future ($+\Delta^t_{j,i} > \alpha$). $sim$ represents the average similarity of a patent to its most similar matched counterparts, and $sim^{past}$, $sim^{present}$, $sim^{future}$ the share of patents in the corresponding time frame weighted

by similarity.

Figure 3: Distribution of similarity by type

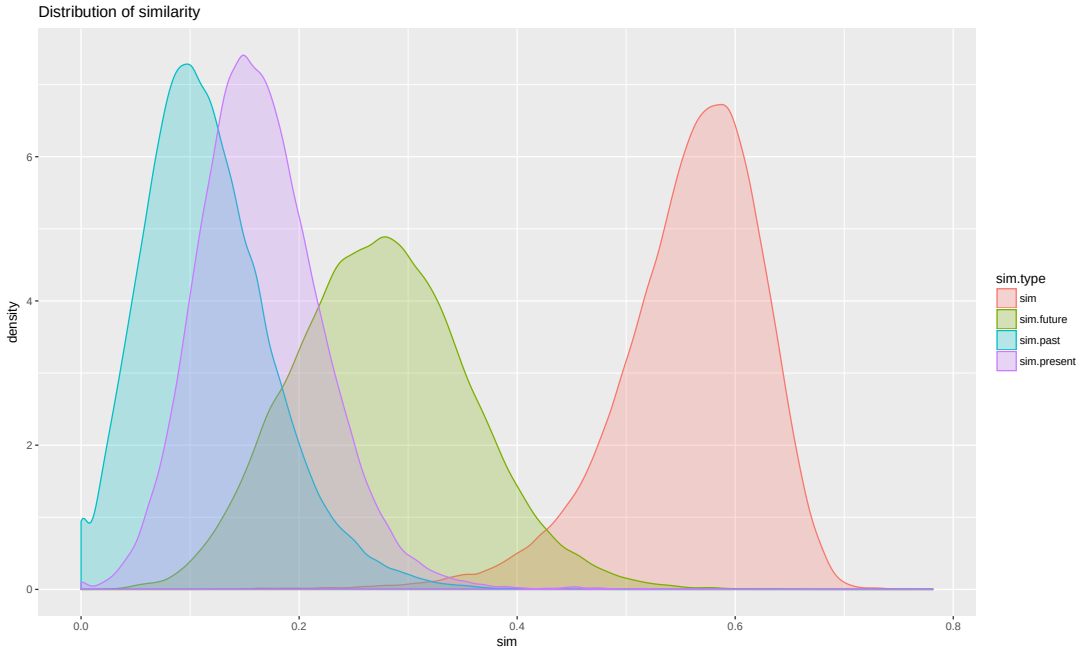
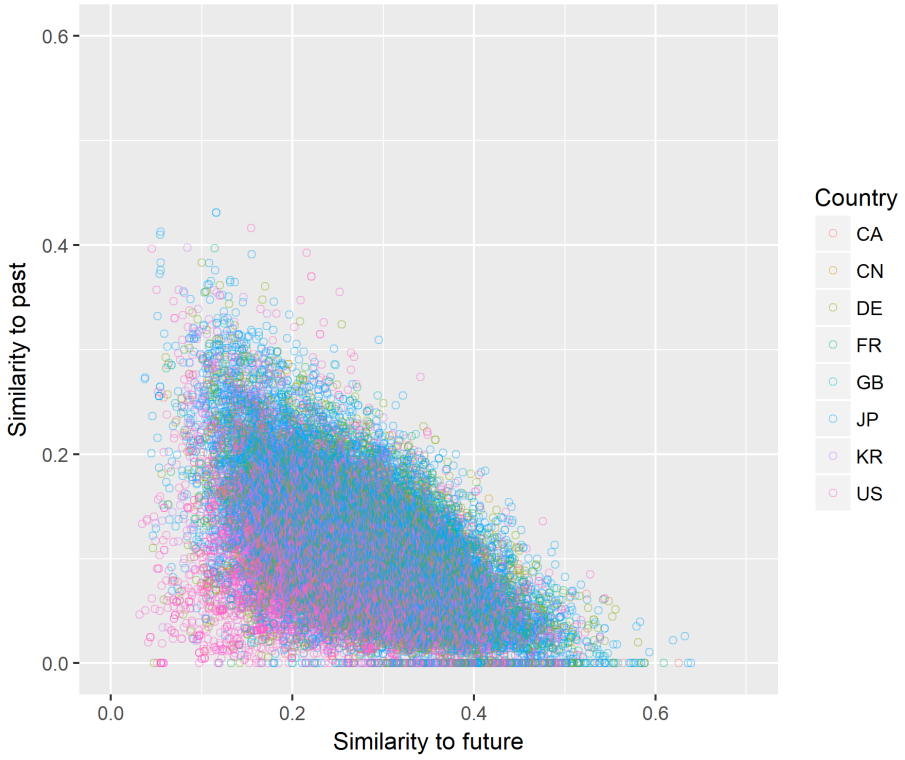



Figure 3 shows the distribution of similarity ($sim$), as well as its decomposition in time ($sim^{past}$, $sim^{present}$, $sim^{future}$). Overall similarity, as expected, appears to be normally distributed, with the highest density slightly above 0.5. Since $sim^{past}$, $sim^{present}$, and $sim^{future}$ represent a decomposition of the overall similarity, their density distribution is shifted to lower values and slightly right skewed, yet still proximately normally distributed.

Additionally, figure 4 plots $sim^{past}$ against $sim^{future}$ on patent level. Even though future and past similarity are somewhat exclusive ($r(sim^{past}, sim^{future}) = -0.54^{***}$), it can be seen that in some cases relatively high values on both temporal similarities can be observed. Furthermore, it can be seen that some patents indeed show no similarity with the past, and high similarity to the future. Such patents are likely candidates for being a "breakthrough invention".

Figure 4: Distribution patent orientation



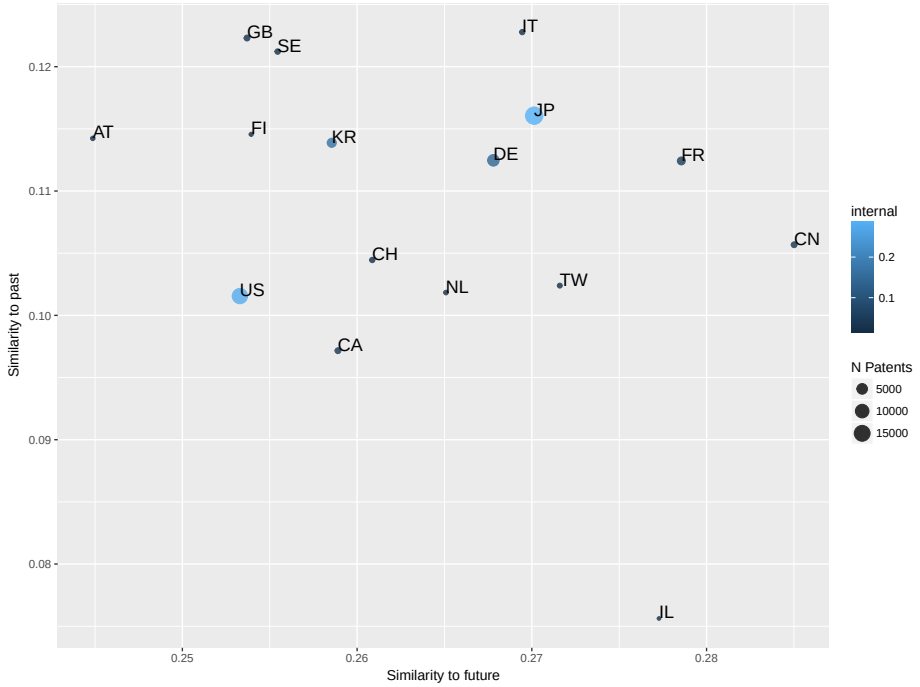

## 4.2 Application: Future impact indicators on different levels of aggregation

After computing temporal similarity scores for every patent, we continue with analyzing the properties and possible insights gained by this indicator on different levels of aggregation.

First, in figure 5, we display the results of an aggregation of temporal similarity on country level, where we plot a country's average similarity scores to the future on the x- and to the past on the y-axis. Conceptually similar to a "Lead-Lag" analysis,[5] the results can be interpreted in several ways: First, a low similarity to the past can be seen as a measure of novelty, the capacity to develop patent applications that are substantially different from common existing technologies. Places in space and points in time of low similarity to the past might also *ex-ante* mark the presence of high

[5] Cf. Nallapati et al. (2011); Ramage et al. (2010); Shi et al. (2010) for several applications of Lead-Lag analysis on academic corpora.

Figure 5: Novelty & Impact on country level

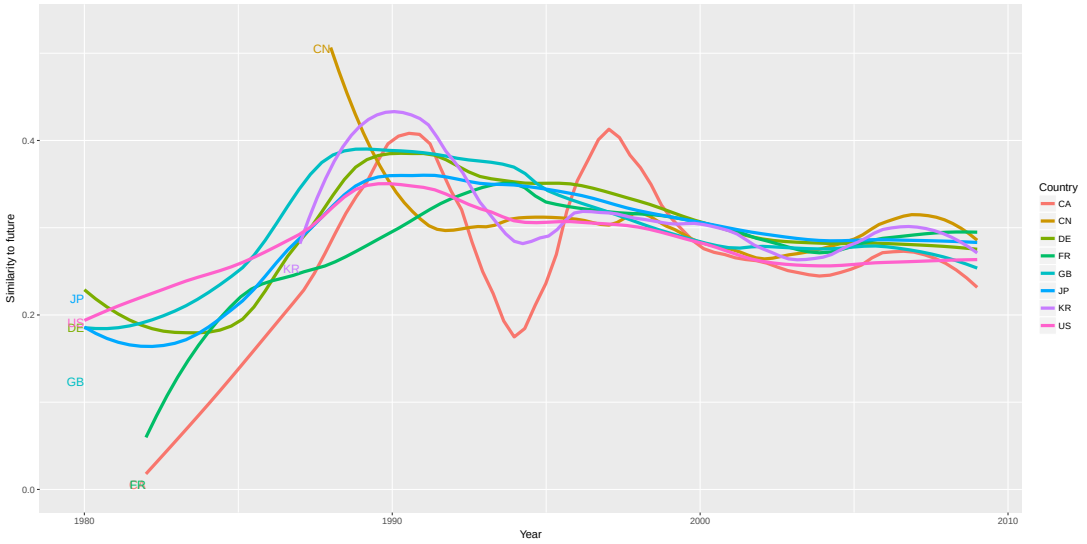



experimental activity, and the emergence of technological discontinuities. Here we see that Israel take the leading position. However, while *ex-ante* novelty is often seen as a necessary condition for the emergence of innovations with high *ex-post* technological or economic impact,[6] novelty *per se* certainly does not represent a sufficient condition thereof. Second, similarity to the future represents the *ex-post* technological impact of a country's patent applications, meaning a large number of similar patents appear typically after the focal patent. The leading country in this measure is China whose patents seem to cover technologies that are particularly directed towards future developments. While some traditional car manufacturing countries such as Germany and Japan show relatively high levels of past but also future orientation is the US more past oriented in its technological developments.

While informative and useful to provide a first glance at the overall technological novelty and impact created in certain countries, a static 30-year aggregation of a tem-

[6]Where at least the necessity of technological novelty can be challenged to be even a necessary condition for an innovation's economic impact, which might be conditional to other firm and market specific characteristics (Simmering and Hain, 2017; Teece, 1986).

Figure 6: Future similarity (impact) on country level over time




poral composition indicator is surely limited in terms of derivable insights. Therefore, in figure 6 we plot the development of $sim^{future}$ over time for the countries with the highest number of considered patent applications. While we do not see a lot of temporal heterogeneity for the "incumbent" countries with the biggest share of patent applications (the "triad" JP, US, DE, with a relatively constant rate), late-comers enter with distinct pattern. For instance, early Canadian electro-mobility patents in the early-$80^s$ show a very low similarity to the future, suggesting them to be mostly incremental adaption of existing technologies from the triad. In contrast, South-Korea's first patents in the mid-$80^s$ show a *au pair* future similarity with most incumbents. In the case of China's entry in the late $80^s$, their future similarity is even above the incumbents, suggesting them tohave entered the field as an innovator rather than an imitator. It is worth noting that the entry of these two Asian countries coincides with a general increase in future similarity, potentially indicating a technological "window of opportunity". This might be related to the diffusion and application of computer technology across all industries and the introduction of advanced microcontrollers to control electric propulsion systems. The heterogeneity, however, vanishes over time leading to a convergence around a constant rate and generally more synchronized

change thereof. This might indicate a successful "catching-up" of the new entrants to the incumbents rate of technological progress, but also a more mature stage of the technology life-cycle with steady incremental progress.

## 5 Discussion and Avenues for Future Research

While the results so far demonstrate the usefulness of a semantic indicator of technological similarity, and give a first glance possible applications, its full potential to a large extend remains unexplored in this paper. In the following, we indicates what needs to be done in order to improve the accuracy of the technological similarity score, to validate its outcome, and apply it to a range of suitable problems.

In a first attempt, we created our indicators within a somewhat homogeneous subset of patents related to the technological field of electro mobility. In a next step, it has to be explored how well our indicators generalize to the whole realm of patents across technological fields. Since our indicators are calculated based on threshold-similarity scores, they are likely to deliver meaningful results in across heterogeneous technologies. However, based on preliminary results we are convinced the presented approach can complement established patent quality indicators by taking patent information into account which has so far mostly been neglected.

We do a first attempt to evaluate the information-richness of the created signature vectors by demonstrating their capacity to predict the patents IPC class. However, further evaluations are also needed here. Related, further improvements can potentially be reached by improving the vector generation process. So far, we have used a large state-of-the-art language model that has been trained by the developers of *spaCy* (Honnibal and Johnson, 2015) using some of the most comprehensive corpora available today. While the abstract language model itself is universal – capturing the features of the English language – the word vectors, which are also used, are general too. Patents, just as legal texts or text from a particular time (e.g. $18^{\text{th}}$ century poetry) use "domain specific" terminology. That is why in such models, better results are expected when using word vectors that have been trained on a corpus of domain specific texts.

In our case, a way forward would be to train additional word vectors on a curated large corpus (several million) of patent data in all related classes, and complement the *en_core_web_lg* model with these vectors. The "quality" of resulting vectors could be again evaluated using a predictive approach. We expect that such improved vectors would not only result in increased prediction accuracy but would also allow to make prediction of patent symbols on a more detailed level.

We up to now provided a first reality test of the proposed indicators by reproducing some stylized facts on the development of electro-mobility technologies. However, a more throughout validation is needed in future work. For instance, a comparison with common indexes for *ex ante* patent novelty (eg., number of backward citations, originality, radicalness index) and *ex post* impact (e.g., backward citations, generality index) and quality (eg., number of claims, patent renewal) are needed.[7] Further evaluations can be done considering our indicator scores of not-granted patents (lack of novelty) and patents receiving popular innovation and invention prices (c.f. Verhoeven et al., 2016). Further evaluations can be done in a case study manner by considering the scores of famous innovations, or the development of well-known technologies.

After a thorough evaluation and validation of the proposed indicators and their properties, we see a range of useful applications to deploy them. First, when following their development over time. the combination of *ex ante* novelty and ex post impact on technology level (see figure 9) has the potential to inform technology life-cycle studies (eg., Gao et al., 2013; Lee et al., 2016) to inform investment decisions, policies, and theory. Likewise, the revision of historical developments of these indicators on country level has could inform studies on technological "catching up" and reveal potential "windows of opportunity" where latecomer countries can enter the technological race (as indicated in figure 6). Similar analysis could be carried out on to investigate the level and dynamics of inventive activity on firm level (see. 8

Another promising avenue of research appears the deployment of these indicators for technology forecasting to enrich existing approaches (eg., Altuntas et al., 2015;

[7] Basically, an evaluation over the whole range of patent quality indicators listed by Squicciarini et al. (2013) appears appropriate.

Kim and Bae, 2017). Of particular interest here is how in combination with other patent characteristics the *ex ante* novelty can contribute to predict a patents *ex post* impact. That can one the one hand be done either in a causal modeling exercise to shed light on the nuanced relationship between novelty and impact, or in a predictive manner, leveraging current advances in machine learning and artificial intelligence.[8] In case such prediction efforts proof feasible and accurate, they can be leveraged for near-real time and granular "nowcasting" and "placecasting" of inventive activity and its quality, similar to recent efforts on entrepreneurial activity (Andrews et al., 2017; Fazio et al., 2016; Guzman and Stern, 2015, 2017, eg.,), or for the rare event prediction of technological breakthroughs.

Further, the dyadic and temporal nature of the proposed similarity measures naturally suggests their deployment for network analysis on different levels of aggregation. For instance, directed similarity networks from patents to similar ones in the future can be used to map and analyze knowledge flows between individuals, firms, and countries (see figure 10 for a first application on country level). Such network representations can also contribute to efforts of mapping and understanding technology landscapes, trajectories, and their evolution (eg., Aharonson and Schilling, 2016; Jurowetzki and Hain, 2014; Mina et al., 2007; Verspagen, 2007).

Lastly, the proposed indicators also have the potential to generalize across other domains of knowledge production and novelty creation with similar data-structure, such as academic publications and research grants.

## 6 Conclusion

In this paper, we developed a novel set of indicators for *ex ante* patent novelty and *ex post* technological impact. We do so by utilizing the rich information on the technological characteristics and features of patent contained in its abstract to create a vector capturing the technological signature of each patent. In a further step, we create

[8] For a general discussion on causal vs. predictive modeling, and the potential of novel machine learning and artificial intelligence methods in economic research, consider Hain and Jurowetzki (ming)

our novelty and impact by analyzing the temporal distribution of patent similarity. We illustrate first results at the example of patents related to the technological field of electro-mobility, and point towards promising avenues for future research to improve, validate, and deploy the proposed set of indicators across a range of applications.

We thereby contribute to the existing body of knowledge in several ways. First, we are among the first to deploy state-of-the-art NLP techniques to the characterization, classification, and valuation of patents. While there recently has been increasing effort in leveraging textual data in patent research, this has for the most part been limited to keywords and term frequencies. In contrast, we apply a linguistic and semantic informed vector space model to create a high dimensional signature vector of the patent, containing rich informations on the technologies embedded in the patent. We provide a first evaluation of the information-richness of this vector to predict the patent's IPC class. Second, we use these vectors to compute dyadic similarities between patents, where we avoid the computation of a full similarity matrix by applying fast approximate nearest neighborhood matching techniques, and thereby provide an efficient method pipeline which scales near-linear and therefore is appropriate for massive datasets. Third, we overcome shortcomings in commonly applied measures of technological relatedness, novelty, and quality by a temporally unbound alternative which is not dependent of explicit indications of knowledge sources in form of citations by the applicants.

We hope the proposed set of indicators pass further validation tests, generalize well across different technology classes and related corpora, and stimulate further research.

## References


Abbas, A., Zhang, L., and Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information*, 37:3–13.

Aharonson, B. S. and Schilling, M. A. (2016). Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. *Research Policy*, 45(1):81–96.

Ahuja, G. and Lampert, C. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic management journal*, 22(6-7):521–543.

Altuntas, S., Dereli, T., and Kusiak, A. (2015). Forecasting technology success based on patent data. *Technological Forecasting and Social Change*, 96:202–214.

Andrews, R. J., Fazio, C., Guzman, J., and Stern, S. . (2017). The startup cartography project: A map of entrepreneurial quality and quantity in the united states across time and location. MIT Working Paper.

Basberg, B. L. (1987). Patents and the measurement of technological change: a survey of the literature. *Research policy*, 16(2-4):131–141.

Bernhardsson, E. (2017). Annoy: Approximate nearest neighbors in c++/python optimized for memory usage and loading/saving to disk.

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., and Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Dahlin, K. B. and Behrens, D. M. (2005). When is an invention really radical?: Defining and measuring technological radicalness. *research policy*, 34(5):717–737.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Deffke, U. (2013). Electric mobility - rethinking the car. Federal Ministry of Education and Research (BMBF), Department for Electronic Systems and Electric Mobility. Web Page.

Don, S. and Min, D. (2016). Feature selection for automatic categorization of patent documents. *Indian Journal of Science and Technology*, 9(37).

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *Bell Communications Research*, pages 281–285.

Ernst, H. (2001). Patent applications and subsequent changes of performance: evidence from time-series cross-section analyses on the firm level. *Research Policy*, 30(1):143–157.

Fall, C. J., TÃűrcsvÃąri, A., Benzineb, K., and Karetka, G. (2003). Automated categorization in the international patent classification. In *Acm Sigir Forum*, volume 37, pages 10–25. ACM.

Fazio, C., Guzman, J., Murray, F., and Stern, S. (2016). A new view of the skew: Quantitative assessment of the quality of american entrepreneurship. MIT Innovation Initiative Paper.

Gambardella, A., Harhoff, D., and Verspagen, B. (2008). The value of European patents. *European Management Review*, 5(2):69–84.

Gao, L., Porter, A. L., Wang, J., Fang, S., Zhang, X., Ma, T., Wang, W., and Huang, L. (2013). Technology life cycle analysis method based on patent documents. *Technological Forecasting and Social Change*, 80(3):398–407.

Gerken, J. M. and Moehrle, M. G. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. *Scientometrics*, 91(3):645–670.

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, 28(4):1661–1707.

Guzman, J. and Stern, S. (2015). Where is silicon valley? *Science*, 347(6222):606–609.

Guzman, J. and Stern, S. (2017). Nowcasting and placecasting entrepreneurial quality and performance. In Haltiwanger, J., Hurst, E., Miranda, J., and Schoar, A., editors, *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, chapter 2. University of Chicago Press.

Hagedoorn, J. and Cloodt, M. (2003). Measuring innovative performance: is there an advantage in using multiple indicators? *Research Policy*, 32(8):1365–1379.

Hain, D. S. and Jurowetzki, R. (forthcoming). The potentials of machine learning and big data in entrepreneurship research – the liaison of econometrics and data science. In Cowling, M. and Saridakis, G., editors, *Handbook of Quantitative Research Methods in Entrepreneurship*. Edward Elgar Publishing.

Hall, B. H. and Harhoff, D. (2012). Recent research on the economics of patents. *Annu. Rev. Econ.*, 4(1):541–565.

Hall, B. H., Jaffe, A., and Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, 36(1):16–38.

Harhoff, D., Scherer, F. M., and Vopel, K. (2003a). Citations, family size, opposition and the value of patent rights. *Research Policy*, 32(8):1343–1363.

Harhoff, D., Scherer, F. M., and Vopel, K. (2003b). Citations, family size, opposition and the value of patent rights. *Research policy*, 32(8):1343–1363.

Harhoff, D. and Wagner, S. (2009). The duration of patent examination at the european patent office. *Management Science*, 55(12):1969–1984.

Hirschey, M. and Richardson, V. J. (2004). Are scientific indicators of patent quality useful to investors? *Journal of Empirical Finance*, 11(1):91–107.

Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Jurowetzki, R. and Hain, D. S. (2014). Mapping the (r-) evolution of technological fields – a semantic network approach. In Aiello, L. M. and McFarland, D., editors, *Social Informatics*, volume 8851 of *Lecture Notes in Computer Science*, pages 359–383. Springer International Publishing.

Kim, G. and Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change*, 117:228–237.

Kim, M., Park, Y., and Yoon, J. (2016). Generating patent development maps for technology monitoring using semantic patent-topic analysis. *Computers & Industrial Engineering*, 98:289–299.

Lanjouw, J. O. and Schankerman, M. (2001). Characteristics of patent litigation: a window on competition. *RAND journal of economics*, pages 129–151.

Lanjouw, J. O. and Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal*, 114(495):441–465.

Larminie, J. and Lowry, J. (2003). *Electric vehicle technology explained,*.

Lee, C., Kim, J., Kwon, O., and Woo, H.-G. (2016). Stochastic technology life cycle analysis using multiple patent indicators. *Technological Forecasting and Social Change*, 106:53–64.

Lee, S., Yoon, B., and Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, 29(6-7):481–497.

Lerner, J. (1994). The importance of patent scope: an empirical analysis. *The RAND Journal of Economics*, pages 319–333.

Liu, D.-R. and Shih, M.-J. (2011). Hybrid-patent classification based on patent-network analysis. *Journal of the Association for Information Science and Technology*, 62(2):246–256.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mina, A., Ramlogan, R., Tampubolon, G., and Metcalfe, J. S. (2007). Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. *Research Policy*, 36(5):789–806.

Nallapati, R., Shi, X., McFarland, D. A., Leskovec, J., and Jurafsky, D. (2011). Leadlag LDA: Estimating topic specific leads and lags of information outlets. In *ICWSM*.

Narin, F., Hamilton, K. S., and Olivastro, D. (1997). The increasing linkage between us technology and public science. *Research policy*, 26(3):317–330.

Niosi, J. (2005). *Canada's Regional Innovation System: The Science-based Industries*. Number Book, Whole. McGill-Queen's University Press, Quebec.

Pakes, A. and Schankerman, M. (1984). The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources. In *R&D, patents, and productivity*, pages 73–88. University of Chicago Press.

Passing, F. and Moehrle, M. G. (2015). Measuring technological convergence in the field of smart grids: A semantic patent analysis approach using textual corpora of technologies. In *Management of Engineering and Technology (PICMET), 2015 Portland International Conference on*, pages 559–570. IEEE.

Pavitt, K. (1985). Patent statistics as indicators of innovative activities: possibilities and problems. *Scientometrics*, 7(1):77–99.

Pavitt, K. (1988). Uses and abuses of patent statistics. In *Handbook of quantitative studies of science and technology*, pages 509–536. Elsevier.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Pilkington, A. and Dyerson, R. (2006). Innovation in disruptive regulatory environments: A patent study of electric vehicle technology development. *European Journal of Innovation Management*, 9(1):79–91.

Pilkington, A., Dyerson, R., and Tissier, O. (2002). The electric vehicle:: Patent data as indicators of technological development. *World Patent Information*, 24(1):5–12.

Proff, H. and Kilian, D. (2012). Competitiveness of the EU Automotive Industry in Electric Vehicles: Final Report. (Journal, Electronic).

Ramage, D., Manning, C. D., and McFarland, D. A. (2010). Which universities lead and lag? toward university rankings based on scholarly output. In *Proc. of NIPS Workshop on Computational Social Science and the Wisdom of the Crowds*. Citeseer.

Schmookler, J. (1966). *Invention and economic growth*. Harvard Univ. Press, Cambridge, MA.

Schoenmakers, W. and Duysters, G. (2010). The technological origins of radical inventions. *Research Policy*, 39(8):1051–1059.

Shane, S. (2001). Technological opportunities and new firm creation. *Management science*, 47(2):205–220.

Shi, X., Nallapati, R., Leskovec, J., McFarland, D., and Jurafsky, D. (2010). Who leads whom: Topical lead-lag analysis across corpora. In *NIPS Workshop*.

Simmering, P. and Hain, D. (2017). Innovation and imitation strategies in the age of the upgrade–an agent-based simulation model. In *Proceedings of the 10$^{th}$ European Meeting on Applied Evolutionary Economics (EMAEE)*.

Squicciarini, M., Dernis, H., and C, C. (2013). Measuring patent quality: Indicators of technological and economic value.

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research policy*, 15(6):285–305.

Tong, X. and Davidson, F. (1994). Measuring national technological performance with patent claims data. *Research Policy*, 23(2):133–141.

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand Journal of Economics*, pages 172–187.

Trajtenberg, M., Henderson, R., and Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and new technology*, 5(1):19–50.

Tran, T. and Kavuluru, R. (2017). Supervised Approaches to Assign Cooperative Patent Classification (CPC) Codes to Patents. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 22–34. Springer.

Verhoeven, D., Bakker, J., and Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, 45(3):707–723.

Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, 10(01):93–115.

WIPO (2017). Guide to the International Patent Classification.

# 7 Appendix

Figure 7: Preprocessing pipeline

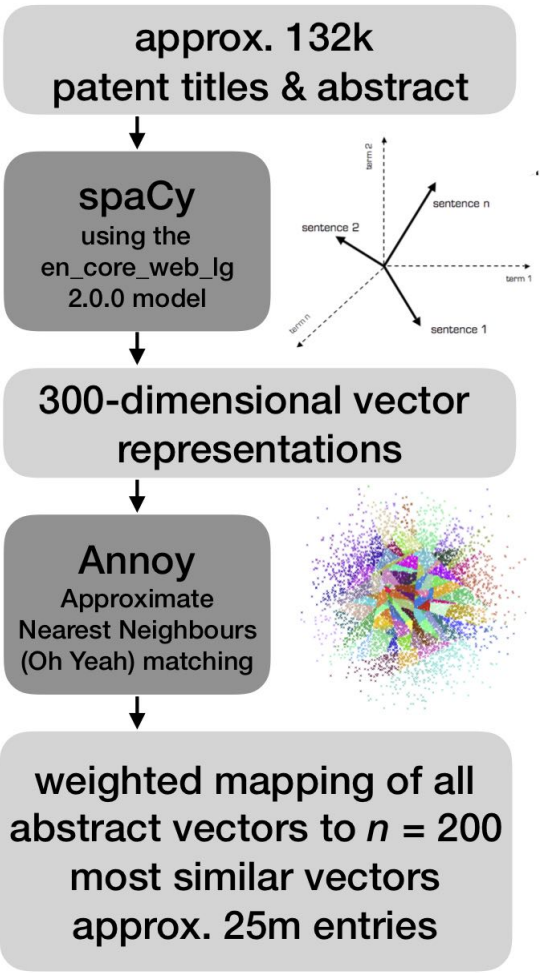


Figure 8: Novelty & Impact on firm level

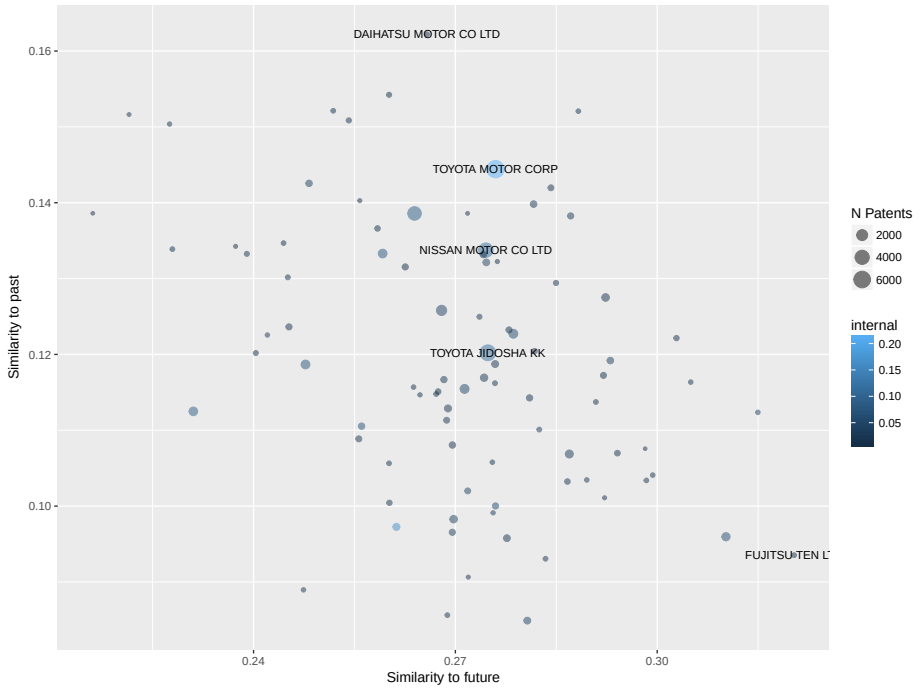



Figure 9: Novelty & Impact on IPC level

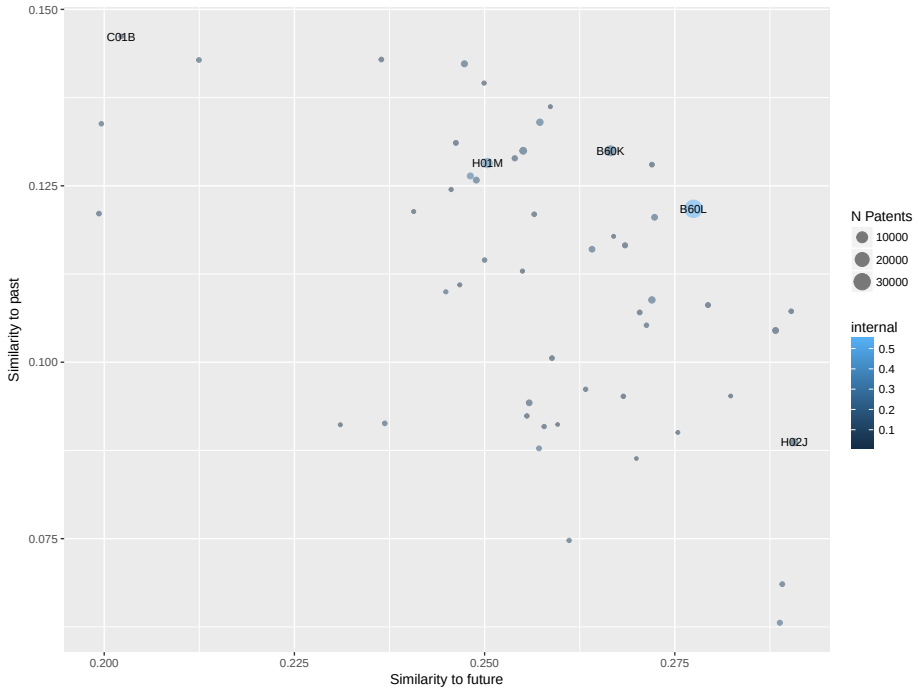
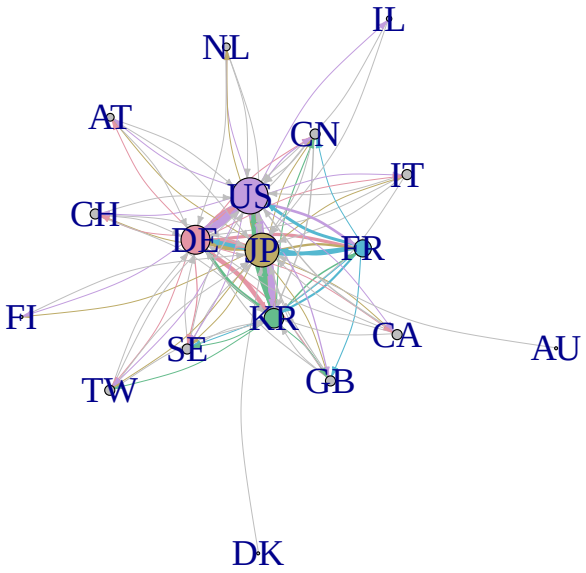


Figure 10: Network: Knowledge flows between countries




*Note:* Static network of knowledge adaption between countries. The directed edges represent the Jaccard-weighted similarity between between ego country $i$ and future patents of alter country $j$. Node size is scaled by a country's out-degree, indicating high influence on adjacent countries technological development.